Validating Dialect Comparison Methods

Wilbert Heeringa[1], John Nerbonne[1], Peter Kleiweg[1]

[1]University of Groningen, Faculty of Arts, Alfa-Informatica,
P.O.Box 716, 9700 AS Groningen, The Netherlands

Abstract: The range of dialectometric methods suggests the need for validation work. We propose a gold standard, based on the consensual classification of a well-studied area. Fidelity to the gold standard is assessed via matrix overlap measures (Rand and Fowlkes/Mallows). Word-based techniques in which varieties are compared to each other directly emerge as superior.

1 Introduction

Séguy (1971) and Goebl (1982, 1984) were among the first to advocate extensive deployment of statistical classification techniques in dialectology. This paper focus on methods which aim at measuring the phonetic distance between varieties at an aggregated level, Hoppenbrouwers and Hoppenbrouwers's (1988), Kessler's (1995) and Nerbonne et al's (1996). The techniques should generalize straightforwardly to lexical distance measures such as Séguy's (1971).

1.1 Comparison Methods

A great number of alternative methods have been proposed for comparing and classifying dialects. Many of the alternatives are refinements of one other, leading to the question which methods are most suitable in general. The present paper examines the performance of a range of methods on a well-understood area, The Netherlands.

We examine dialect distance measurements varying several dimensions.

1. Unit of measurement: word vs. corpus.
   Hoppenbrouwers and Hoppenbrouwers (1988) attain respectable results in holistic measures of corpora. Kessler (1995) and others measure differences per word.
2. Direct comparison vs. comparison via standard language.
3. Phonetic representation: phones vs. phonetic features.
   PHONES are letter-like units which can be described via a small number of phonetic properties, their FEATURES.
4. In feature-based measures: feature system.
   We compare a system developed for measuring the accuracy of phonetic transcription (Vierregge) to a phonologically motivated system (Hoppenbrouwers's)

5. In feature-based measures: distance measure between feature bundles, which can be determined via Euclidean distance, Manhattan (city-block) distance, or via a measure based on Pearson's r.
6. Order sensitivity: Levenshtein distance vs. feature/phone bags. Levenshtein distance counts *on* and *no* as different, while other measures simply count phones or features.
7. In Levenshtein measures: value of insertions and deletions. These values may be determined either with respect to a logical maximum or with respect to maxima in existing material.
8. Sensitivity of measure to frequency (Information Gain Weighting).
9. Representation of diphthongs (complex vowels) as one vs. two segments.

Although not all of the nine dimensions combine with one another, we nonetheless examine 334 combinations. The variety reinforces the need for validation techniques.

2 "Gold Standard" Validation

The leading idea in our validation is that dialectometric methods ought to agree with expert consensus in well-studied cases, which we take to be a "Gold Standard." A gold standard provides therefore a classification of language varieties with which (nearly) all experts agree. When expert dialectologists disagree on a well-studied variety (which often happens near area borders, for example), then that variety cannot be part of the gold standard. In this way the gold standard is incomplete, but it represents consensus.

Our research is based on 104 local varieties, for which the data is taken from the *Reeks Nederlands(ch)e Dialectatlassen* (Blancquaert & Peé (1925–1982)). From the transcriptions in the atlas we chose 100 words as being representative of the dialect. The gold standard was defined with respect to these 104 varieties from the RND, and their phonetic data was input to the distance procedures.

Two authoritative dialect maps were taken as starting points, namely the map of Van Ginneken (1913) which we took from Weijnen (1966), and the map of Daan, found in Daan & Blok (1969). The map of Van Ginneken is based on objective linguistic criteria, together with the judgement of the dialectologist (Goossens (1977)). The map of Daan is based on the "arrow" method. Dialects which are judged (nearly) equal by the speakers are connected by arrows. Connected dialect points are dialect areas, and borders emerge as small strips which are not crossed by arrows.

Both maps are criticized (legitimately or not), but this does not disqualify them from use in our project. We found three levels of comparability:

partitions of three, five and twelve areas. We regard the sum of errors in these divisions as a comprehensive error.

2.1 Distance Validation

We examine various methods from two perspectives derived from distance. The first perspective emphasizes that varieties within the gold-standard groups ought to be relatively close to each other as compared to varieties outside the gold-standard group. We check whether the distances within groups ("within") are much smaller than distances between groups ("without"). We choose this terminology to emphasize the similarity between this calculation and the well-known F distribution. For each group g_k in the gold standard we calculate the ratio between the mean squared within-group distance and the mean squared without-group distance. Let d_{ij} be the distance between dialect i and dialect j:

$$within_{g_k} = \frac{\sum_{i \in g_k} \sum_{j \in g_k} {d_{ij}}^2}{|d_{i \in g_k j \in g_k}|} \tag{1}$$

$$without_{g_k} = \frac{\sum_{i \in g_k} \sum_{j \notin g_k} {d_{ij}}^2}{|d_{i \in g_k j \notin g_k}|} \tag{2}$$

$$F\ ratio = \sum_{k=1}^{c} \frac{within_{g_k}}{without_{g_k}} \tag{3}$$

The last equation determines an overall value for the assignment in a gold standard with c groups. The lower the F-ratio, the better the distance assignment corresponds with the gold standard. We normalize F-ratio by dividing it by c.

A second perspective derived directly from distance is that of discrimination. A distance assignment discriminates well if the (gold standard) groups emerge clearly. To operationalize this idea, we regard a dialect as a point in n-space fixed by its distance to the n dialects in the sample. We then expect gold standard groups to occupy relatively small regions (compared to non groups), with minimal overlap among groups, and thus maximal discrimination. Fisher's Linear Discriminant measures the discrimination between two groups (Schalkoff (1992)). The idea is that group means should differ maximally with respect to group variances. We calculated group discrimination for each dimension. If the gold standard has c groups among n varieties, the total discrimination between all group pairs over all dimensions is:

$$D = \sum_{d=1}^{n} \sum_{i=2}^{c} \sum_{j=1}^{i-1} \frac{(\mu_{di} - \mu_{dj})^2}{\sigma_{di}^2 + \sigma_{dj}^2} \tag{4}$$

where i, j ranges over dialect groups. We normalize D by dividing it by the number of group pairs $\binom{c}{2}$ and dimensions n. Larger values indicate better discrimination.

2.2 Validation via Classification

We clustered distance matrices to obtain classifications. We examined seven clustering techniques: single link, complete link, group average, weighted average, unweighted centroid, weighted centroid and minimum variance, also known as Ward's method (Jain & Dubes (1988)).

In validation, we use a gold standard which is specified as a partition. Therefore the dialect distances we calculated were converted to a partition as well so that the one partition could be compared to the other. From the dendrogram we could derive a partition of k groups, where k is equal to the number of groups in the gold standard ($2 \leq k \leq n-1$). Note that we don't use the most detailed hierarchical information in the dendrogram, only the highest level divisions.

Rand Index

Having two partitions we compare them with each other by using the Rand index (Rand (1971)) which Hubert & Arabie (1985) recommend as "one of the most popular alternatives for comparing partitions..." Given n dialects, suppose we have a partition M, based on the distances of the method we want to validate, and a partition G which is the gold standard. Both partitions consist of k groups. Each of the $\binom{n}{2}$ dialect pairs belongs to one of the following types:

1. M and G assign the dialects to the same group;
2. M and G each assign the dialects to different groups;
3. M assigns the dialects to different groups but G to the same; or
4. M assigns the dialects to the same group, but G to different ones;

1 and 2 are agreements, while 3 and 4 represent disagreements. We construct a matrix, where the rows correspond with the groups of M, and the columns correspond with the groups of G. Let n_{ij} be the number of dialects in group i of M and group j of G, $n_{i\cdot}$ the total number of dialects in group i of M and $n_{\cdot j}$ the total number of dialects in group

j, and g be the number of groups in the partitions under comparison, $g = |G| = |M|$. The number of agreements R_k is:

$$\binom{n}{2} + 2\sum_{i=1}^{g}\sum_{j=1}^{g}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{g}\binom{n_{i\cdot}}{2} + \sum_{j=1}^{g}\binom{n_{\cdot j}}{2}\right] \tag{5}$$

The number of disagreements is equal to $\binom{n}{2} - agreements$. The error rate is expressed as the probability of a disagreement: $disagreements/\binom{n}{2}$.

Fowlkes and Mallows Index

An other method for comparing partitions was developed by Fowlkes & Mallows (1983). A brief description can be found in Hubert & Arabie (1985). Their measure of association B_k is defined as:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\binom{n_{ij}}{2} / \sqrt{\sum_{i=1}^{g}\binom{n_{i\cdot}}{2} * \sum_{j=1}^{g}\binom{n_{\cdot j}}{2}} \tag{6}$$

If each group in M perfectly matches with a group in G, B_k is 1. If each group in M is equally distributed over all groups in G, B_k is 0.

3 Consistency

In the case of word-based methods, we can check on consistency using Cronbach's α, which is derived from the inter-item correlation of the words. Each of the more than 5×10^3 dialect pairs is assigned a separate distance based on each of the 100 words. We calculate the correlation r between words $w1$ and $w2$ in the usual way, for each pair of words. This is summarized in the average correlation $\bar{r}$. Cronbach's α is calculated as follows:

$$\alpha = \frac{n_w \bar{r}}{1 + (n_w - 1)\bar{r}}, \text{ where } n_w \text{ is the number of words} \tag{7}$$

4 Which Validation Methods?

Both the F ratio and Fisher's Linear Discriminant tend to select techniques which maximize the contrast between groups. The F emphasizes group coherence with respect to contrast to other groups (reflected in dendrograms), and Fisher's Linear Discriminant emphasizes contrast, but relative to the variance of the distance measures. Only the second

validation technique was able to identify methods which were "linguistically successful"—ones which led to successful classifications after clustering according to our gold standard. The methods which scored optimally according to the F ratio emphasized contrast and seemed to underassess group-internal diversity. When we examined dendrograms produced by these methods, there tended to be virtually no group-internal distance, something which the best methods consistently recognize. We found this surprising, and had expected the F ratio to be more useful.

Fisher's Linear Discriminant is much more successful than the F ratio; in general, methods which score well here led to clusterings which were close to the gold standard. Details were often wrong, however, particularly in cases with larger numbers of groups.

Both the Rand index and the Fowlkes and Mallows index work very well, and, furthermore, they tended to agree most on the best methods. On the basis of 104 dialects and summarized over all three levels of the gold standard, both indexes judges the same method as the best.

These results are perhaps not surprising if one recalls that the gold standard is essentially nominal. The distance measures are of course metric, which is why metric criteria such as Fisher's Linear Discriminant apply at all, but the nominal matrix-overlap criteria (Rand and Fowlkes/Mallows) measure most directly what the gold standard wants: a classification of dialects.

5 Which Dialectometric Methods?

Table 1 shows results for the Rand index, the Fowlkes and Mallows index and Cronbach's α. We average results' scores in order to address questions of choice separately. See the caption for a summary of results.

The same method came out as optimal according to both the Rand index and the Fowlkes/Mallows: Levenshtein using unweighted Vieregge features with 1-segment diphthongs, in which varieties were compared directly (rather than through standard Dutch), and in which feature vector distance was assayed via $1 - r$. The value of insertions and deletions was determined with reference to a null vector, which might be interpreted as silence, and which also worked best on average among Levenshtein methods.

Literatur

BLANCQUAERT, E. and PEÉ W. (1925–1982): *Reeks Nederlands(ch)e Dialectatlassen*, De Sikkel, Antwerpen.

DAAN, J. and BLOK D. P. (1969): *Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde*, Noord-Hollandsche Uitgevers Maatschappij, Amsterdam.

	Rand	F & M	Cronbach
Comparison method			
corpus frequency	0.64	1.51	
frequency per word	0.54	1.81	0.94
Levenshtein indel from corpus	0.55	1.79	0.93
Levenshtein indel theoretically	0.52	1.79	0.91
Levenshtein indel variable	0.60	1.73	0.95
Phone representation			
phones	0.56	1.79	0.90
features Hoppenbrouwers redundant	0.55	1.76	0.88
features Vieregge	0.53	1.77	0.88
features Hoppenbrouwers not redundant	0.54	1.78	0.88
Representation diphthongs			
as two segments	0.55	1.76	0.88
as one segment	0.54	1.78	0.88
Comparison feature histograms/bundles			
Manhattan	0.54	1.77	0.88
Euclidean	0.55	1.77	0.89
'Pearson'	0.54	1.76	0.88
Frequency weighting of features			
no weighting	0.54	1.78	0.88
weighting	0.55	1.76	0.88
Direct or indirectly comparison			
direct	0.55	1.78	0.93
indirect	0.51	1.82	0.83
Cluster methods			
single link	0.78	1.50	
complete link	0.34	1.99	
group average	0.36	1.94	
weighted average	0.39	1.91	
unweighted centroid	0.89	1.40	
weighted centroid	0.85	1.43	
minimum variance	0.20	2.23	
Totals			
worse	0.98	1.20	0.79
best	0.07	2.56	0.96

Table 1: Mean method scores using Rand, Fowlkes/Mallows (F & M) and Cronbach's α (Cronbach). All word methods are very reliable with 100 words (Cronbach). Minimum variance clustering, direct comparison (without reference to standard language) and word-based methods are clearly superior. Frequency Weighting leads to slightly worse results; two-segment representation of diphthongs is slightly better, and feature-comparison schemes are roughly equivalent. There appear to be dependencies among these choices — the best don't simply use all the best (average) choices. See text for further details.

FOWLKES, E. B. and MALLOWS, C. L. (1983): 'A Method for Comparing Two Hierarchical Clusterings', *Journal of the American Statistical Association* **78**, 553–569.

GOEBL, H. (1982): *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*, Österreichische Akademie der Wissenschaften, Wien.

GOEBL, H. (1984): *Dialektometrische Studien. Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterial aus AIS und ALF.* 3 volumes, Niemeyer, Tübingen.

GOOSSENS, J. (1977): *Inleiding tot de Nederlandse Dialectologie*, Wolters-Noordhoff, Groningen.

HOPPENBROUWERS, C. and HOPPENBROUWERS, G. (1988): 'De featurefrequentiemethode en de classificatie van nederlandse dialecten', *TABU: Bulletin voor taalwetenschap* **18**(2), 51–92.

HUBERT, L. J. and ARABIE, P. (1985): 'Comparing Partitions', *Journal of Classification* **2**, 193–218.

JAIN, A. K. and DUBES, R. C. (1988): *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, New Jersey.

KESSLER, B. (1995): Computational Dialectology in Irish Gaelic, *in* 'Proceedings of the European Association for Computational Linguistics', EACL, Dublin, pp. 60–67.

NERBONNE, J. et al. (1996): Phonetic Distance between Dutch Dialects, *in* G. Durieux, W. Daelemans and S. Gillis, eds, 'CLIN VI, Papers from the sixth CLIN meeting', University of Antwerp, Center for Dutch Language and Speech, Antwerp, pp. 185–202.

RAND, W. M. (1971): 'Objective Criterion for Evaluation of Clustering Methods', *Journal of the American Statistical Association* **66**, 846–850.

SCHALKOFF, R. (1992): *Pattern Recognition: Statistical, Structural and Neural Approaches*, John Wiley & Sons, Inc., New York.

SÉGUY, J. (1971): 'La relation entre la distance spatiale et la distance lexicale', *Revue de Linguistique Romane* 35, 335-357.

VAN GINNEKEN, J. (1913): *Handboek der Nederlandsche taal. Deel I: De sociologische structuur der Nederlandsche taal*, Nijmegen.

VIEREGGE, W. H. (1987): Basic Aspects of Phonetic Segmental Transcription, *in* A. Almeida and A. Braun, eds, 'Probleme der phonetischen Transkription, *Zeitschrift für Dialektologie und Linguistik*, Beihefte', Franz Steiner Verlag Wiesbaden GMBH, pp. 5–55.

WEIJNEN, A. (1966): *Nederlandse dialectkunde*, Van Gorcum, Assen.